ML Hw6 report

1. **(2%) 試說明 hw6_best.sh 攻擊的方法，包括使用的 proxy model、方法、參數等。此方法和 FGSM 的差異為何？如何影響你的結果？請完整討論。(依內容完整度給分)**

   在經過多次嘗試後，我選擇 densenet121 作為我的 proxy model ，而設定的 ϵ 為 0.1，使用的 loss function 為nll_loss，而使用的攻擊方法依然為FGSM。用這個攻擊方法在black box 得到的攻擊成功率為0.92，L-infinity為5.625。

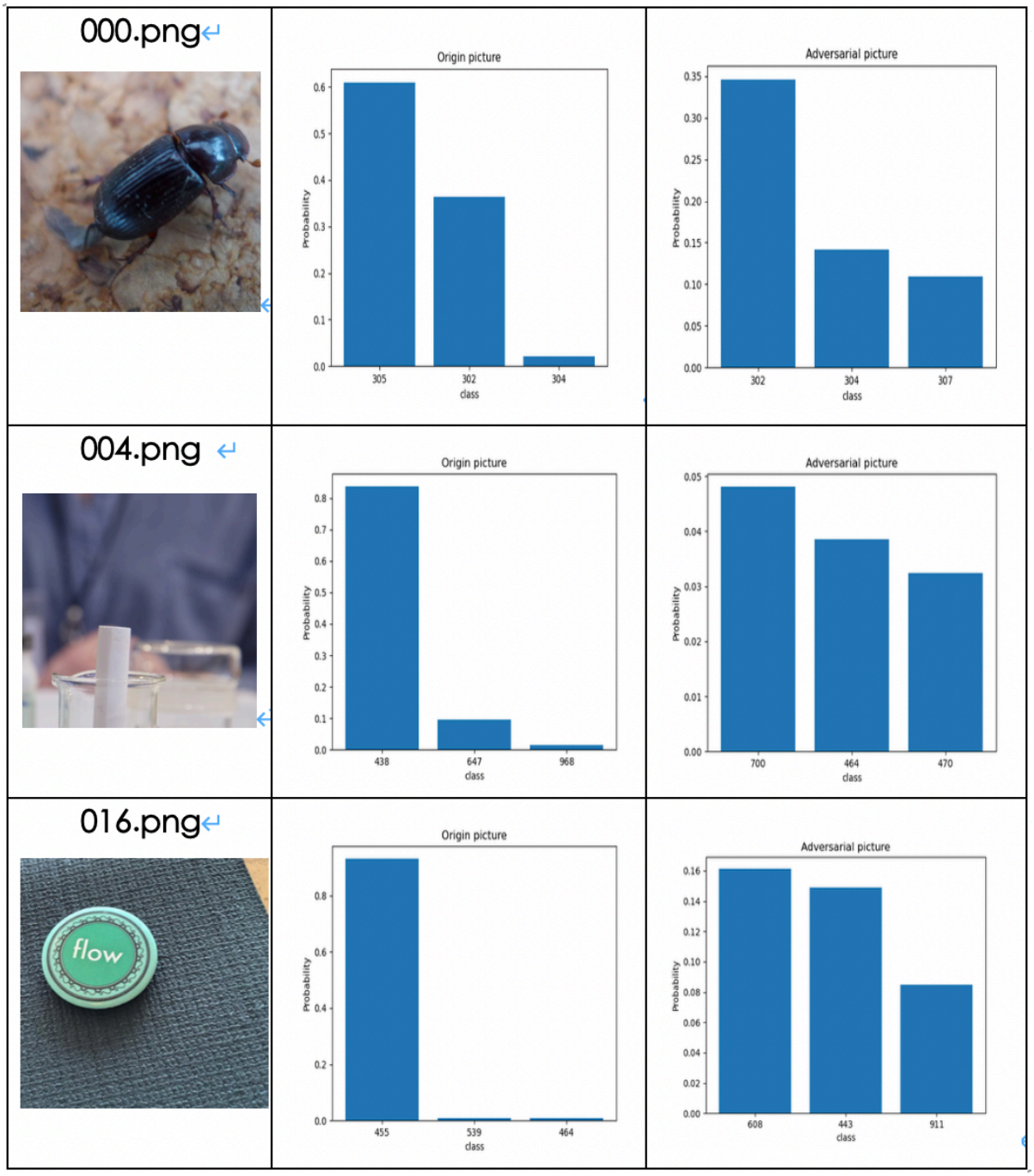
   我也有嘗試過 Basic Iterative Method的方法，將更新的learning rate設為0.001，並使用 Adam 作為gradient descent 的optimizer，並且將原本圖片和新圖片的差別項 $\mu(X_N - X_{N+1})^2$ 加在 loss function 中（μ為調整用的參數，設為1000，以達到限制$d(x, x_0) \leq \epsilon$ 的條件）。使用這個方式可能是我試的不夠多，因此在測試上效果就沒有那麼好（後來無法在judge boi 上嘗試新的 submittion，所以無法得知對black box attack的效果如何）

2. **(1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。**

$\epsilon = 0.1$	**attack success rate**	**L-infinity**
Vgg16	0.265	5.325
Vgg19	0.265	5.35
Densenet121	0.920	5.625
Densenet169	0.400	5.575
Resnet50	0.275	5.425
Resnet101	0.465	5.325

   如上表所示，使用的proxy model為Densenet121時攻擊的成功率最高，而理論也顯示雖然不同的模型也能attack成功，但使用與black box同一種proxy model應可使成功率最高，因此背後的 black box 應該最有可能是 Densenet121。

3. **(1%) 請以 hw6_best.sh 的方法，visualize 任意三張圖片攻擊前後的機率圖 (分別取前三高的機率)。**



可以由上面三張照片的攻擊中得知，原本每一個照片都有一個很明確的class（此處的機率是由softmax函數計算），但是加上攻擊的微小雜訊後，分佈變得相當平緩，不會有任何一項是特別突出的。且除了第一張照片以外，原本機率最高的三個class在經過攻擊之後都沒有出現在前三名，這代表此model的確是被攻擊成功。

4. **1. (2%) 請將你產生出來的 adversarial img，以任一種 smoothing 的方式實作被動防禦 (passive defense)，觀察是否有效降低模型的誤判的比例。請說明你的方法，附上你防禦前後的 success rate，並簡要說明你的觀察。另外也請討論此防禦對原始圖片會有什麼影響。**

   我的方法是實作被動式防禦中使用filter的方式，在網路上查到guassian filter的實作或許對這樣的攻擊是有效的，而在上傳judge 之後可以發現success rate 從防禦前的0.92 降到了0.605。這種 filter 可以達到濾除雜訊、低通、模糊化圖片的效果，因此可能就改變了原本圖片在被攻擊的那個方向（gradiant）的資訊，進而達到好的防禦效果。